The Software Alliance
BSA

バイアスに挑む：
AIの信頼性構築に向けたBSAのフレームワーク

www.bsa.org

目次

はじめに

人工知能(AI)の研究開発が飛躍的に進歩したことにより、テクノロジーによって世界が様変わりする過程において人々の期待が急速に変化しています。いつかAIがあらゆる業界に影響を与えるようになるだろうという予想は、ビジネスにおいて急速に現実になりつつあります。金融サービスから医療などのさまざまな分野で、顧客体験の向上や競争力強化のため、また以前ならば解決が難しかった問題への対処のために、AIがますます活用されています。たとえばAIにより、医学研究者は衰弱症状が発生する何年も前に早期アルツハイマー病を診断できるようになっています。[1]また、環境学者が膨大なデータセットを分析する上でも役立っており、絶滅の危機にある生息地の保護や、マラウイでの象の密猟阻止に対する取り組みの効果を、容易に追跡することができるようになっています。[2]

本レポートで使用する「人工知能」とは、機械学習アルゴリズムを組み込んだシステムを指します。このアルゴリズムは大量の訓練データを分析して、相関関係やパターンを特定することができます。また、その他のメタデータを特定して、将来のデータインプットに基づいて予想や推奨を行うためのモデルを開発する際に、使用することもできます。たとえば、開発者は機械学習を使用して「Seeing AI」というアプリを作成しました。このアプリは、画像中のオブジェクトに音声による説明を加えるもので、目の不自由な人が外出する際の手助けになっています。[3]このアプリのユーザーがスマートフォンを使用して画像を撮影すると、Seeing AIが画像の中にあるものについて説明します。画像内のオブジェクトを識別できるコンピューターの視覚モデルを開発するため、一般的に入手可能な、樹木、道路標識、風景、動物など身の回りのオブジェクトが写った何百万もの画像から得たデータを用いて、システムの訓練が行われました。ユーザーが新しい画像をアプリに入れると、Seeing AIはその画像と訓練データで学習したパターンや相関とを比較して、画像中にどのようなオブジェクトがあるかを推定します。

一方、あらゆる業界でAIの利用が急増していることから、テクノロジーの設計や使用について、また社会にもたらしうる潜在的リスクを考慮して運用するためにどのような対策を行うかについて、問題が提起されています。

リスクの高い意思決定をする上で先端テクノロジーを使用することは、機会を得るのと同時にリスクを取ることにもなります。一方で金融機関でのAI導入については、データに基づく、人間によるバイアスの影響を受けにくい意思決定方法を取り入れることで、差別が減り、公平性が高められる可能性があります。[4]たとえばAIを使用することで、貸し手が通常、従来の信用報告書から得ているデータに比べて、より多くのデータを分析できるようになります。その結果、歴史的に疎外されてきたコミュニティの人々が貸し付けを受けたり、住宅を確保する機会が増えると考えられます。同時に研究者たちは、AIシステムの設計、開発および導入における欠陥が、既存の社会的バイアスを長期化(さらには悪化)させる可能性があると警鐘を鳴らしています。[5]

そのため、AIバイアスリスクを特定して軽減する仕組みの開発は、産学官の専門家にとっての重点分野として注目されています。ここ数年間に行われた膨大な研究により、組織のベスト・プラクティス、ガバナンス上の保護措置、およびAIのライフサイクル全体にわたってバイアスリスクを管理する上で役立つテクニカルツールが幅広く特定されました。AIモデルの静的評価により、AIシステムが現場に導入された際に生じうる潜在的な問題がすべてカバーされるわけではありません。そのため専門家の間では、AIバイアスによるリスクを軽減するためには、エンドユーザーが継続的なモニタリングを行ってシステムが意図された通りに動いていることを確認するなど、ライフサイクルを通した対策を取る必要があるとの考えで意見が一致しています。

本レポートは、*AIバイアスリスク管理のフレームワーク*を定めるものであり、組織はこのフレームワークを使用することで影響評価（インパクト・アセスメント）を行い、AIシステムのライフサイクル全体を通して発生する可能性のある潜在的なバイアスリスクを特定し、軽減することができます。データプライバシーの影響評価と同様に、AIの影響評価は確実性のための重要な仕組みとなり、説明責任を強化するものです。またこれにより、ハイリスクなAIシステムが設計、開発、試験の後に導入され、危害のリスクを軽減するための十分な防御対策も講じられているとの信頼性を高めることができます。AIの影響評価は、透明性を保つ仕組みでもあります。これにより、AIシステムの設計、開発、導入に関わる多くの潜在的な利害関係者がリスクについて話し合い、リスク軽減のための責任を明確に理解することができます。

AIの影響評価を行うためのプロセスの確立に加えて、バイアスリスク管理のフレームワークには次のような事項が含まれます。

- AIリスク管理プログラムを効果的に実施し、支援する上で必要となる、企業ガバナンスの重要な仕組み、プロセス、保護措置を提示します。
- AIシステムのライフサイクルを通して発生するおそれのある、AIに固有のバイアスリスクを軽減するための、既存のベスト・プラクティス、テクニカルツール、リソースについて確認します。

このフレームワークは公平性、透明性および説明責任を促進するリスク管理プロセスを通じて、組織がAIシステムへの信頼性を高めるために使用できる柔軟なツールとなることを目的としています。

AIバイアスとは?

本レポートで使用する「AIバイアス」という用語は、特定の属性を持つ人々に対してシステム的かつ不当に、不都合、不公平または有害な結果をもたらすようなAIシステムを指します。

機械学習の目標は、過去の事例から一般化された規則を導き出すモデルを構築し、将来のデータインプットを予測することです。たとえば、植物を識別するために設計された画像認識システムは、多様な種類の植物が一つひとつ写った大量の画像を読み込むことで訓練されると考えられます。このシステムは、それぞれの品種の画像に共通する一般的な規則(葉模様など)を探して、それに基づきモデルを構築することで、新しいデータインプット(ユーザーが提供した画像)に、識別訓練された品種が含まれているかを評価できるようになります。つまり機械学習は、将来のデータインプットを予測するために過去のデータに基づいた一般化を行うものです。しかし、AIを使用して人間の行動をモデル化する場合、意図しないバイアスについての懸念がまったく異なる次元で生じます。AIが業務プロセスに搭載され、人々の生活に多大な影響を持つ現在、「バイアスを持った」システムにより、歴史的に疎外されてきたコミュニティの人々がシステム上、不利な立場に置かれるリスクがあります。AIバイアスが発生する可能性があるのは、正確性に欠けるシステム、または人種、性自認、性的指向、年齢、宗教、障害の有無などの(これらに限らない)センシティブな特性に基づき、人々が不利になるような判断をするシステムです。

AIバイアスの原因と種類

設計

AIバイアスは、AIのライフサイクルのさまざまな段階で組み込まれる可能性があります。[6]AIシステムの構想と設計の初期段階で行われた決定が、次のようなバイアスをもたらす可能性があります。

- **問題形成バイアス：**場合によっては、提案されたAIシステムの根本となる基本的な仮定に本質的なバイアスがあり、どのような形でも一般展開には適さないことがあります。

EXAMPLES

2016年、上海交通大学の研究者たちが、顔認証システムによって犯罪行為を予測するAIシステムの訓練に向けた研究について論文[7]を発表し、大変な議論を呼びました。警察が撮影した大量の犯人の顔写真を用いてシステムを訓練した結果、システムが人間の顔の造りを分析するだけで、90%に近い精度で犯罪行為を予想することができたと、研究者らは主張しました。当然のことながら、この論文はすぐに厳しい非難の対象となり、論評者たちは当然ながら、このモデルは人の外見から犯罪行為を予測できるとの、深刻に憂慮すべき（そして因果関係として裏付けのない）仮定に基づいていると指摘しました。[8]

システムが実際に予測しようとしているものに対して、AIシステムの目標変数が間違っている、あるいは単純すぎるプロキシである場合にも、問題形成バイアスが生じる可能性があります。たとえば病院では、とあるAIシステムが広く使用されており、急を要する治療が必要な可能性を予測することで患者の重症度判定[9]をしていますが、このAIシステムが、より軽症の白人患者の治療をシステム上優先し、より重症のマイノリティ患者が不利になっていたことが、2019年の研究で明らかになりました。この例でバイアスが発生したのは、システムが「治療の必要性」を予測する際に、患者の治療の必要性を示す実際的なデータの代わりとして、簡単に利用することができる「治療費」の履歴データを使用したためです。マイノリティ患者はこれまで治療を受ける機会が少なかったため、マイノリティ患者が治療を受ける現時点での必要性ではなく、「医療費」に基づいて予測することで、残念ながら間違った状況が描き出されることになるのです。そして、危険なまでにバイアスがかかった結果となる可能性があります。

- **歴史的バイアス：**AIシステムの訓練に使用されるデータに含まれる歴史的バイアスが長期化するリスクがあります。

EXAMPLE

英国のある医学部が、入学者候補にふさわしい学生を特定するため、システムの構築に着手しました。そのシステムは、過去に合格した学生に関するデータを使用して訓練されました。しかし、その大学の過去の合格者決定では、マイノリティや女性が、属性以外の条件がその他の出願者と同等であっても、システム上不利な扱いを受けていたことが明らかになりました。歴史的バイアスを含むデータを使用してモデルを訓練したため、この医学部はうかつにも、同様のバイアスがかかった合格判定パターンを繰り返すようなシステムを作ってしまったのです。[10]

- **サンプリング バイアス：**システムの訓練に使われるデータに、使用される母集団についての間違った説明が含まれていると、訓練データで過小評価されていたコミュニティではシステムが効果的に機能しないおそれがあります。このような現象は、十分な量の説明データがすぐに入手できない場合、また特定の母集団をシステム上、過大（過小）評価するような方法でデータが選択、収集された場合に、よく発生します。

EXAMPLES

Joy BuolamwiniとTimnit Gebruが行なった草分け的研究では、白人男性の顔に偏って集められたデータセットに基づいて顔認識システムを訓練すると、そのシステムが肌の色が濃い女性の顔を認識する際には、正確性がかなり低くなることが証明されました。[11]

サンプリングバイアスは、データ収集方法が原因となって発生する場合もあります。そのよい例として、修理が必要な道路の穴を自動的に検出し、報告するシステムを構築しようとしたボストン市の試みが挙げられます。このプログラムの初期のバージョンでは、「StreetBump」というスマートフォンアプリのユーザーから提供されたデータに大きく依存していました。そのため、スマートフォンやデータプランを購入できる人が多く住む、裕福な地域から大量の報告がなされました。サンプリングバイアスのために、貧しい地域の道路穴がデータセット上では少なく示されることとなり、システムによる修理予算の配分の結果、そのような地域の住人が不平等に扱われるリスクが生じました。[12]

- **ラベリングバイアス：**多くのAIシステムでは、学習アルゴリズムが将来のデータインプットを分類する際に使用できるパターンや相関を見つけられるようにするため、訓練データに「ラベル」を付ける必要があります。訓練データセットにラベリングする過程で、主観的な決定がなされる場合があり、それがAIシステムに人間のバイアスを組み込むきっかけとなる可能性があります。

EXAMPLE

ImageNetは、1400万を超える画像を分類してラベルを付けたデータベースで、これによりAIの研究者が視覚認識システムを訓練しています。ImageNetは、AIによる物体認識の最先端技術の発展にとって重要なツールです。しかし近年の研究では、人々の画像でシステムを訓練するのに使用する際、データベースの分類およびラベリング方法によって、バイアスによる重大リスクがどのように発生するかが注目されています。*Excavating AI（AIの発掘）*の中で[13]、Kate CrawfordとTrevor Paglenは、ImageNetに含まれる人々の画像に付けられた分類やデータラベルには、性、人種、障害の有無、年齢に基づくバイアスが含まれており、それらを訓練データとして使用するあらゆるAIシステムに同様のバイアスが伝播するおそれがあることを明らかにしました。たとえば、ImageNetのデータを使用して訓練されたAIシステムでは、黒人対象者の画像が「不法行為者」または「犯罪者」として分類される傾向が高かったのです。[14]

開発

必要なデータが収集されたら、開発チームはデータを整理、処理、標準化して、モデルの訓練および検証に使用できるようにする必要があります。また開発者は、機械学習手法を選択するか、市販のモデルの中から、使用するデータや解決しようとしている問題の性質にふさわしいものを採用しなければなりません。これには、さまざまな手法を用いてたくさんのモデルを構築し、それらの中から最も有効なモデルを選択するという場合もあるでしょう。[15]通常開発チームは、モデルを機能させるために、データパラメータについても選択する必要があります。たとえば、数値スコアを表すデータは、しきい値を設けて「はい」または「いいえ」の回答に変換される場合があります一つまり、Xまたはそれ以上のスコアは「はい」として再指定され、そのしきい値以下のスコアは「いいえ」と指定されます。開発段階では、次のようなバイアスが生じる可能性があります。

- **プロキシバイアス：**訓練中にモデルが比較する入力変数（「特徴」など）を選択するプロセスは、バイアスをもたらすもう1つの重要な決定ポイントです。センシティブな属性データが除外されていても、システムが依存する特徴がこれらの特性（プロキシと呼ばれる）に深く関連していると、バイアスが生じることがあります。

EXAMPLE

一見問題のない特徴を使用していても、センシティブな属性と相関があるために、プロキシバイアスが生じる可能性があります。たとえば、ある人がMacか、それ以外のノートパソコンを所有しているかという情報から、ローンの返済可能性が予測できる場合があることが、研究によって明らかになっています。[16]したがって金融機関は、ローン申込者から候補者を選別するAIシステムを構築する際、このような変数の組み込みを検討する可能性があります。しかし、Macパソコンの所有は人種と密接に関係しているため、この特徴を含めることでも、プロキシバイアスによる大きなリスクが発生します。その結果、そのような特徴を含めることにより、人種に深く相関しているが、実際の信用リスクには関係のない特徴に基づいて、申込者をシステム上冷遇するシステムができあがる可能性があります。

- **集計バイアス：**「汎用的な」モデルを使用し、重要な変数が見落とされると、そのシステム性能は主要なサブグループに対してしか活かすことができないものとなるおそれがあります。モデルがシステムの正確性に重大な影響を与えるサブグループ間の根本的な違いを考慮していないと、集計バイアスが生じる可能性があります。平均値や集計値では、特異な事例が見落とされるおそれがあるのです。それどころか、集団についての値を集計したモデルにより、同じ集団内のサブグループの状態について、異なる動き、あるいは逆の動きが正確に予測される場合もあります。この現象は、シンプソンのパラドックスと呼ばれています。

EXAMPLE

集計バイアスリスクは特に、医療現場で顕著に現れます。診断や治療では、病気がさまざまな人種、民族の人々に影響を与える際の特有の傾向を考慮する必要があります。たとえば、糖尿病による合併症のリスクは民族によって大きく異なるため、糖尿病に関連するリスクを予測するAIシステムは、このような違いを考慮しなければ、一部の患者に対して精度が下がる可能性があります。[17]

導入、モニタリング、および反復

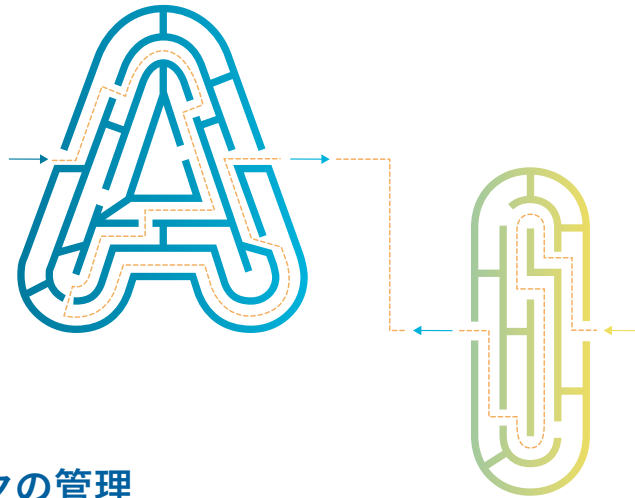
AIシステムが、モデルの訓練に使用されるデータと異なる現実のシナリオに出会うことは避けられません。そのため、導入前に綿密な検証や試験が行われたシステムでも、運用が始められると性能が低下する可能性があります。したがって、AIシステムのライフサイクル全体を通じて、評価や査定を継続的に行うことが重要です。

- **導入バイアス：**システムが導入された後にも、さまざまな形でバイアスが発生する可能性があります。たとえば、AIシステムの訓練または評価に使用されたデータが、システム導入後に処理の対象となる集団のデータと著しく異なる場合、モデルが予定通りに機能することができなくなります。訓練時にモデルが過剰に調整されたこと（予測モデルが訓練データについての大量な詳細情報を学習したため、その他のデータについて正確な推論ができない）、あるいはコンセプトの傾向（目標変数と訓練データの相関における変化により性能が悪化した）のどちらかが原因で、訓練に使われたデータの範囲外の事象について、モデルが信頼性のある推論ができない場合に、導入バイアスが生じる可能性があります。

- **誤用バイアス：**また、1つの目的のために構築されたAIシステムや特徴が予期しなかった方法や意図されなかった方法で使用された場合にも、導入バイアスが発生する可能性があります。

AIリスク管理の必要性

リスク管理とは？

リスク管理とは、リスクを特定し、潜在的影響を緩和する方法論を確立することで、設計によりシステムの信頼性を確保するプロセスです。リスク管理プロセスは、サイバーセキュリティやプライバシーなどの分野において特に重要です。これらの分野では、テクノロジーの急速な進化と、脅威内容の激しい変化が組み合わさることで、従来の「コンプライアンス」ベースのアプローチが効果を発揮しなくなっています。すぐに陳腐化するような、変化のない規定要件リストに照らして製品やサービスを評価するのではなく、リスク管理では製品やサービスのライフサイクル全体を通じてリスクを軽減するために、コンプライアンス責任を開発過程に統合することを目指しています。効果的なリスク管理は、製品の設計、開発、導入の重要な局面における組織の開発チームとコンプライアンス担当者の協力を高めるガバナンス・フレームワークに基づいています。

バイアスのリスクの管理

AIシステムを開発・使用する組織は対策を講じて、バイアスの発生により、ある人の属性についての特徴に基づいて、不都合または有害な結果が不当にもたらされることがないようにする必要があります。このようなバイアスから生じうる被害を効果的に防御するにはリスク管理のアプローチが必要であり、その理由は次の通りです。

「バイアス」と「公平性」は状況に応じて異なる

システムが「公平」に動作しているかを評価する方法について一般的な合意がないため、AIシステムからのバイアスを取り除くことはできません。実際、Arvind Narayanan教授による有名な説明にもある通り、システムが公平に動作しているかを評価するための定義[18]（つまり数字上の基準）は少なくとも21種類あるため、AIシステムがそのすべてを同時に満たすことは*不可能*なのです。公平性についての普遍的な定義はありません。そのため、開発者は自らが作成しているシステムの内容を評価して、バイアスを特定するためにはどの数値基準が最も適切かを判断し、もたらされうるリスクを軽減する必要があります。

バイアス軽減対策には、代償が伴う可能性もある

ある集団のバイアスを軽減する対策が、別の集団でバイアスを増加させ、システム全体の精度が低下する可能性もあります。[19]リスク管理により、そのような代償に状況に応じたやり方で対応する手段が得られます。

導入後にバイアスが発生する可能性もある

システムが導入前に十分に検討されていても、誤用されたり、属性が訓練や試験データにおけるそれと異なるような環境に導入されたりした場合、バイアスのある結果が生じる可能性があります。

効果的なリスク管理の基盤

リスク管理の目的は、AIシステムのライフサイクル全体で発生しうる潜在的リスクを特定し、軽減するために、繰り返し使用できるプロセスを確立することです。包括的なリスク管理プログラムには、次の2つの重要な要素があります：

1
組織のリスク管理機能を下支えするための**ガバナンス・フレームワーク**。

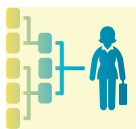
2
影響評価を行ってリスクを特定し軽減するための、数値評価が可能なプロセス。

ガバナンス・フレームワーク

効率的なAIリスク管理を行うには、システムのライフサイクル全体を通してリスクを特定、軽減、記録する際に拠りどころとなる方針、プロセス、および担当者が定められたガバナンス・フレームワークを基礎とする必要があります。このようなガバナンス・フレームワークの目的は、組織内の部署全体(製品開発、コンプライアンス、マーケティング、営業、経営幹部など)に対して、AIシステムの設計、開発および導入において効果的なリスク管理を進める上で、それぞれの部署が担う役割や責任について理解を広めることです。リスク管理ガバナンス・フレームワークの主な特徴は次のとおりです。

方針とプロセス

ガバナンス・フレームワークの中心となるのは、組織のリスク管理に対するアプローチを定めた各種の公式な方針です。これらの方針において、組織のリスク管理目標、その目標を達成するために従う手順、そしてコンプライアンスの評価の拠りどころとなる基準を定義する必要があります。

- **目標:**AIリスク管理は、組織がその基本理念に基づいてAIを開発・使用することを目的に、組織におけるさまざまなリスク管理機能の状況に即して行われる必要があります。そのために、ガバナンス・フレームワークで、基本理念を脅かす可能性のあるリスクをどのように管理するかを定める必要があります。
- **プロセス:**ガバナンス・フレームワークでは、リスクを特定し、リスクの重大さを評価し、AIライフサイクルの各段階でリスクを軽減するための、プロセスと手順を確立する必要があります。
- **評価システム:**組織が方針や手順が規定通りに実施されているかを評価する際に使用する数値基準やベンチマークなどの決まりについて、ガバナンス・フレームワークにて規定する必要があります。
- **定期的なレビュー:**AIの能力が今後も成熟し、テクノロジーが新たな用途で使われる中で、組織はAIガバナンス・フレームワークを定期的に見直し、更新して、目的に合った状態に維持し、進化するリスク内容に対応できるようにすることが重要です。

経営者による監視:AI開発者とAI導入担当者は、経営者による十分な監視によって強化されたガバナンス・フレームワークを維持する必要があります。経営幹部はガバナンス・フレームワークの方針内容を作成して承認するだけでなく、企業におけるAI製品開発ライフサイクルを監視して、積極的にその役割を果たさなければなりません。結果的に人々に悪影響を及ぼす可能性のあるハイリスクなシステムについては、会社の経営陣が責任を持って「実行/実行しない」の意思決定を行う必要があります。

担当者、役割および責任

リスク管理の効果が出るかは、AIのライフサイクルを通して意思決定を取り仕切るための、職能上の枠を超えた専門家グループを設立できるかによって決まります。組織の規模や、開発または導入するシステムの性質によっては、リスク管理の責任において複数部門のスタッフの関与が必要になることもあります。このためガバナンス・フレームワークでは、AIリスク管理に関連する役割と責任を担う組織内の担当者を定め、報告ライン、権限、および必要な専門知識をはっきりと位置づける必要があります。役割や責任を割り当てる際、組織は独立性、能力、影響力、多様性の優先順位を考える必要があります。

- **独立性:**複数レベルで独自のレビューができるように担当者を配置することで、リスク管理を最も効果的に行うことができます。たとえば、リスク管理の責任は、次のような複数チームに分割することができます。
  - **製品開発チーム:**AI製品・サービスの設計開発に携わるエンジニア、データ・サイエンティスト、各種専門家。
  - **コンプライアンスチーム:**ハイリスクなAIシステムにおける影響評価の開発など、AI開発において方針や実務が遵守されているかを監視する法律、コンプライアンスおよび各分野の専門家、またデータ専門家を含め、多種多様に構成されるチーム。
  - **ガバナンスチーム:**組織のAIガバナンス・フレームワークとリスク管理プロセスの効果的な監視を、計画、維持、確認する役割を担うチーム。理想的には、経営幹部が主導する。
- **能力、情報提供および影響力:**リスク管理の責任を担う担当者は、ガバナンスの役割を果たすため、十分なトレーニングや情報提供を受ける必要があります。また、担当者が権限を与えられることや、リスクに対処したり、上長に報告したりする際の意思決定を行う上で効果的なインセンティブを与えられることも同様に重要です。たとえば、組織は明確な報告ルートを定めてリスク管理担当者が経営幹部の意思決定者と連携できるようにし、重要なリスク領域や決定について経営幹部レベルでの見通しが得られるようにしなければなりません。

多様性:AIシステムは社会的、技術的性質を持つため、システムの開発と監視に関わるチーム内の多様性に優先順位を付けることが非常に重要になります。開発および監視プロセスは、チームメンバーがさまざまな視点や背景を持っている場合に最も効果を発揮します。そのような視点や背景は、AIシステムの影響を受けたり、AIシステムを利用したりするユーザーのニーズや懸念を推定する上で有益です。「アルゴリズム開発では、倫理感や政治的思想など、開発者が自覚しない思い込みが暗黙のうちにコード化される」ため、組織においては、さまざまな実体験が投影されたチームを構成することや、AIの設計・開発プロセスのライフサイクルの全期間に、これまで軽視されてきた考え方を取り入れることが重要です。[20]組織が多様性に欠けている場合は、外部の利害関係者に相談の上、特にシステムの影響によって過小評価されてきた集団からのフィードバックを得る必要があります。

影響評価

AIリスク管理を効果的に行うには、一般の人々に大きく影響を与えうるシステムの影響評価を行うための、盤石なプロセスを踏む必要があります。影響評価は、環境保護からデータ保護に至るまでさまざまな分野で幅広く使用されています。影響評価は説明責任の仕組みであり、システムが一般社会に及ぼしうる潜在的リスクを考慮して設計されていることを説明することで、信頼性が高まるのです。つまり、影響評価の目的は、システムがもたらしうるリスクを特定し、システムが与えうる被害の規模を数値化し、そして許容できるレベルまでリスクを軽減するために取られた対応を記録することです。

評価の対象であるシステムの性質や、それによって生じうる被害の内容に合わせて、影響評価プロセスを調整する必要があります。極めてローリスクなシステムには(たとえば文書に使われるフォントの種類を予測するシステム)、完全な影響評価を行う必要がない場合もあります。しかし、システムに本質的なリスクがあり、一般社会に物的被害をもたらす可能性がある場合には、完全な影響評価が行われる必要があります。AIは驚くほどさまざまなアプリケーションに搭載できるため、リスクを特定して軽減するための「万能な」アプローチはありません。その代わりに、影響評価プロセスを調整することで、AIシステムの性質と、それによって生じうる本質的なリスクや物的被害の可能性に備える必要があります。システムに物的被害を発生させる本質的なリスクがあるかを判断する上で、利害関係者は次の点を考慮する必要があります。

- **人々への潜在的な影響:**貸し付けを受けたり、住宅を確保する機会など、AIシステムが意思決定プロセスで使われ、その結果人々に間接的な影響が及ぶような場合でも、影響評価は同じく重要です。
- **システムの状況と目的:**影響評価の必要性と、適切な実施範囲を判断するには、最初にAIシステムの性質と使用環境の評価を行うとよいでしょう。発生しうる被害の深刻度や頻度が高い領域(医療、輸送、金融など)で使われるハイリスクなAIシステムにとって、影響評価は特に重要になります。
- **人間による監視レベル:**どの程度AIシステムが完全に自動化されるかも、AIシステムの本質的なリスクに影響を与える可能性があります。高度に熟練した専門家に提言を行うように設計されたシステムは、同様の完全に自動化されたシステムよりも固有のリスクが少ないと考えられます。もちろん、人間による判断をシステムに組み込むだけで、AIシステムからリスクがなくなるわけではありません。むしろ、人間とコンピューターの相互作用の性質を総合的に精査し、人間の監視によってAIシステムの本質的なリスクをどの程度軽減できるかを見極める必要があります。
- **データの種類:**システムの訓練に使用されるデータの性質によって、システムの本質的なリスクが明らかになる場合もあります。たとえば、使用される訓練データが人間の性格や行動に関するものである場合、それはシステムのバイアスをより慎重に監視しなければならないというサインです。

AIバイアスリスク管理のフレームワーク

以下ではAIバイアスリスク管理のフレームワークの概要を説明します。このフレームワークは、AIバイアスの潜在的リスクを伴うシステムについて、組織が影響評価を行う際の参考となるように設計されています。このフレームワークは、AIシステムのライフサイクルを通して発生するバイアスの原因を特定するプロセスを定めるだけでなく、リスクを軽減するためのベスト・プラクティスを提示するものです。

このフレームワークは、確実性に基づく説明責任のための仕組みであり、AI開発者およびAI導入組織は次のような目的で使用することができます。

- **内部プロセスの指針：**AI開発者とAI導入者がこのフレームワークを使うことで、内部プロセスにおける役割、責任そして想定について、組織立てて定めることができます。
- **トレーニング、意識向上および教育：** AI開発者とAI導入者は、このフレームワークを使うことで、AIシステムの開発と使用に携わる従業員向けの社内トレーニングや教育プログラムを作成することができます。さらにこのフレームワークは、組織のAIバイアスリスク管理へのアプローチについて経営幹部を教育するための有用なツールとなります。
- **保証および説明責任：** AI開発者とAI導入者は、このフレームワークに基づいてそれぞれの役割や責任について話し合い、調整することができ、システムのライフサイクルを通して、AIリスクを管理することができます。
- **ベンダーとの関係：**AI導入者はこのフレームワークを指針として、AIリスクが十分に検討されたことを確認しながら、購買を決定したり、ベンダーとの契約を作成したりできます。
- **信頼と自信：**AI開発者が、製品の特徴や、その製品がAIバイアスリスクを軽減するアプローチについて、一般の人々に伝えたい場合に、このフレームワークは、倫理的なAIシステムを構築する取り組みについて、組織が一般の人々に説明するのに役立ちます。
- **インシデントへの対応：**突然のインシデントが発生した際、このフレームワークが定めるプロセスと参照資料が監査証跡となり、AI開発者やAI導入者がシステムのパフォーマンス低下または不具合の原因を見つけることができます。

AIライフサイクルにおける各フェーズ

このフレームワークはAIライフサイクルの各フェーズに沿って体系化されており、AIシステムを構築し、使用する際に繰り返し行う、重要な手順を説明するものです。

設計フェーズ

- **プロジェクトの概念：**AI設計の初期段階では、システムが解決しようとする「問題」を特定して定義し、モデルによってその目標をどのように達成するかを計画します。このフェーズでは、設計チームがシステムの目的と構造を定義します。システムの性質に応じて、設計チームはシステムが予測する目標変数を特定します。この例としては、あるフィットネスアプリがユーザーの心拍数を分析して異常がないかモニターし、その人に脳卒中または心臓病（つまり目標変数）のリスクがあるかを予測するような場合が挙げられます。システム設計プロセスの初期段階においては、AIの使用が現在のプロジェクトにとって適切か否かを特定することが、バイアスリスク管理のフレームワークの目標です。潜在的リスクは次のとおりです。
  - **問題形成バイアス：**目標変数には、本質的なバイアスまたは正しくない仮定が含まれる場合があり、危険なバイアスが長期化するおそれがあります。場合によっては、提案されたAIシステムの根本となる基本的な仮定に本質的なバイアスがあり、どのような形でも一般展開には適さないことがあります。
- **データ取得：**システムの目標を定義したら、開発者は、データのコーパスを構築する必要があります。このコーパスを使用してモデルがパターンを認識できるよう訓練し、将来のデータインプットについて予測ができるようにするのです。この訓練データがさまざまな形で、思いがけずAIシステムにバイアスを生み出す可能性があります。潜在的リスクは次のとおりです。
  - **歴史的バイアス：**システムを訓練するデータそのものに歴史的バイアスが含まれる場合、不平等性がさらに定着するおそれがあります。
  - **サンプリングバイアス：**AIシステムの訓練に使用されるデータが、システムが対象とする集団についての典型的なデータではない場合にも、バイアスリスクが発生します。典型的でないデータで訓練されたAIシステムは、過大（過小）評価されている階層の人について予測を行う際、効果的に機能しない可能性があります。
  - **ラベリングバイアス：**多くのAIシステムでは、訓練データにラベルを付けて、どのようなパターンを探すかを判断できるようにする必要があります。訓練データセットにラベリングするプロセスは、AIシステムにバイアスを組み込むきっかけとなる可能性があります。

開発フェーズ

- **データ準備とモデルの定義：**AIライフサイクルの次のステップでは、モデルの訓練に備えて、データを準備します。開発チームはこのプロセスにおいて、訓練データにおける変数（つまり「特徴」）を整理し、標準化し、特定します。アルゴリズムはパターンや相関を探す際に、その訓練データを確認し、それを将来の予測を行うための基本規則とします。開発チームはそれに加え、システムを強化するアルゴリズムモデルのタイプ（直線回帰、ロジスティック回帰、ディープニューラルネットワークなど）を選択するなどして、システムの基盤となる構成を組み立てなければなりません。[21]データの準備が整い、アルゴリズムが選択された後、開発チームはシステムを訓練して、将来のデータインプットについて予測ができる機能モデルを作成します。潜在的リスクは次のとおりです。
  - **プロキシバイアス：**訓練データで特徴を選択し、モデリング方法を選択するプロセスでは、モデルの目標変数について予測をする際、どの変数が関連していると見なすかを、人間が判断します。このような判断により、保護されたクラスのプロキシとしての役割を果たす変数を基にするなど、システムに思いがけずバイアスが組み込まれる可能性があります。
  - **集計バイアス：**モデルがシステムの正確性に重大な影響を与えるサブグループ間の根本的な違いを考慮していないと、集計バイアスが生じる可能性があります。「汎用的な」モデルを使用し、重要な変数が見落とされると、そのシステム性能は主要なサブグループに対してしか活かすことができないものとなるおそれがあります。
- **モデルの検証、試験および修正：**モデルの訓練が完了したら、検証によってモデルが意図したとおりに動作しているかを確認し、試験を行ってシステムから得られた結果に予定外のバイアスが含まれていないことを実証する必要があります。検証と試験の結果によっては、容認できないバイアスによるリスクを軽減するため、モデルを修正する必要が生じる可能性もあります。

導入フェーズ

- **導入と使用：**導入前に、AI 開発者はシステムを評価して、設計および開発の初期段階で特定されたリスクが、会社のガバナンスポリシーに対応する方法で十分に軽減されているかどうかを判断する必要があります。特定されたリスクがシステムの誤用によって発生する可能性がある場合、AI開発者は、製品性質（誤用リスクを軽減するユーザーインターフェイスなど）を統合してリスクを軽減し、リスクを悪化させる可能性のある使用（エンドユーザーライセンス契約など）を禁止することで、リスクを抑制する必要があります。また、AI導入担当者に、自身の影響評価を行うための十分な文書を提供します。

  AI システムを使用する前に、AI導入者はAI開発者が提供する文書を確認し、システムが独自のAI ガバナンスポリシーに対応しているかどうかを評価し、導入に関連するリスク管理責任が明確に割り当てられているかどうかを判断する必要があります。AI開発者は導入後のリスク管理責任の一部に対処できますが、AI導入者はシステム性能を監視し、それがリスクプロファイルと整合性のある方法で動作しているかどうか評価する責任を負うことがよくあります。潜在的リスクは次のとおりです。

– **導入バイアス：**AIシステムは、時間の静的な瞬間を表し、一貫性のある正確なモデルの予測能力を損なう可能性のある「ノイズ」を除外するデータでトレーニングされます。実際の環境での導入時に、AI システムは開発環境やテスト環境とは異なる条件に必然的に遭遇します。さらに、現実世界は時間の経過とともに変化するため、データ変数間の関係が進化するにつれて、モデルが表すその時のスナップショットの正確性はどうしても低下することがあります。導入されたAIシステムの入力データが訓練データと大きく異なる場合、システムが「ドリフト」して、バイアスのリスクを悪化させるようにモデルの性能が低下する可能性があります。例えば、AIシステムが特定の国で使用するように設計（およびテスト）されている場合、人口構成が大幅に異なる国で使用されると、システムの性能が良くないことがあります。

– **誤用バイアス：**AIシステムを設計された条件とは大きく異なる環境に導入したり、意図された使用事例と合致しない目的で導入したりすると、バイアスのリスクが悪化する可能性があります。

フレームワーク構造

フレームワークは、システムのライフサイクル全体にわたるAIバイアスのリスクを特定し軽減するためのベストプラクティスを特定します。それは次のように構成されています。

- **機能：**AIリスク管理の基本的な活動を最高レベルで示し、影響評価とリスク軽減のベストプラクティスの間で分割します。

- **カテゴリー：** AIライフサイクルの各フェーズで機能を実行するために必要な活動とプロセスを設定します。つまり、カテゴリーは、影響評価を実行するための手順を設定し、関連するリスクの管理に使用できる対応するリスク軽減のベストプラクティスを特定します。

- **診断ステートメント：** カテゴリーに対して実行する必要がある個別の行動を定義します。これは、各カテゴリーの成果を達成するために役立つ結果のセットを提供します。

- **実装に関するコメント：** 診断ステートメントで説明されている結果を達成するための追加情報を提供します。

- **ツールと資料：** 関係者がAIのライフサイクルの各フェーズに関連するバイアスリスクを軽減するために使用できる、さまざまな外部ガイダンスとツールキットを特定します。フレームワークで特定された特定のツールと資料は、すべてを網羅しているわけではなく、情報提供のみを目的として強調されています。

関係者の役割と責任

フレームワークは、AIシステムの本質的に動的な性質を反映しているので、システムの設計、開発、導入の多様な側面で役割を果たす可能性のある一連の関係者を考慮しています。AIの開発または導入のモデルは一つではないため、理論上は、フレームワークの多くのリスク管理機能には、役割を割り当てたり特定の責任を負わせたりすることができません。しかし、一般的には、システムのライフサイクルを通じたAIリスク管理の特定の側面について、様々な責任を負う三種類の関係者が存在します。

- **AI開発者：**AI開発者は、AIシステムの設計と開発を担当する組織です。
- **AI導入者：**AI導入者は、AIシステムを採用して使用する組織です。（法人が独自のシステムを開発している場合は、それはAI開発者でもAI導入者でもあります）。
- **AIエンドユーザー：**AIエンドユーザーは、AIシステムの使用を監督する責任がある個人で、多くの場合AI導入者の従業員です。

これらの関係者間でのリスク管理責任の分配は、多くの場合、AIシステムの開発および導入モデルに依存しています。

AIの開発および導入モデルのスペクトル

関係者間のリスク管理責任の適切な分配は、開発されるAIシステムの性質、およびどの当事者が基盤となるモデルのトレーニングの目的と方法を決定するかによって異なります。例えば、次のようになります。

- **ユニバーサル、静的モデル：**AI開発者は、すべての顧客（AI導入者）に、事前にトレーニングを受けた静的なモデルを提供します。
  - AI開発者は、モデルリスク管理のほとんどの側面について責任を負います。
- **カスタマイズ可能モデル：**AI開発者は、事前にトレーニングを受けたモデルをAI導入者に提供します。AI導入者は、独自のデータを使用してモデルをカスタマイズおよび/または再トレーニングできます。
  - リスク管理は、AI開発者とAI導入者の間で責任が共有されます。
- **特注モデル：**AI開発者は、AI導入者のデータを使用して、AI導入者に代わって特注AIモデルをトレーニングします。
  - リスク管理は、AI開発者とAI導入者の間で責任を共有しますが、AI導入者により多くの義務が課せられます。

BSA AIバイアスリスク管理フレームワーク

設計

機能	カテゴリー	診断ステートメント	実装に関するコメント
プロジェクトの概念			
影響評価	目標と仮定を特定して文書化する。	システムの意図と目的を文書化する。	• システムの目的は何か。つまり、どのような「問題」を解決するのか。 • システムの対象ユーザーは誰か。 • システムはどこでどのように使用されるか。 • どのような誤用が考えられるか。
		モデルの意図する効果を明確に定義する。	モデルが何を予測、分類、推奨、ランク付け、または発見するように意図されているか。
		意図する使用の事例とシステムが導入される文脈を明確に定義する。	
	公平性を評価するための数値基準を選択して文書化する。	AIシステムのバイアスを評価するための基準として使用する、「公平性」についての数値基準を特定する。	「公平性」という概念は非常に主観的であり、評価に用いることができる数値基準は多数ある。公平性の数値基準をすべて同時に満たすことは不可能なため、開発中のAIシステムの性質に最も適しており、適用法の要件に一致する基準を選択する必要がある。AIライフサイクルの後半の段階における情報提供のため、公平性の数値基準を選択および/または除外する為に用いる根拠を文書化することが重要である。
	関係者の影響を文書化する	システムの影響を受ける可能性のある関係者グループを特定する。	関係者グループには、AI導入者、AIエンドユーザー、影響を受ける個人（AIシステムとやり取りしたり、AIシステムによって影響を受けたりする可能性のある一般のメンバー）が含まれる。
		各利害関係者グループについて、システムの意図する使用と合理的に予見可能な誤使用の両方を考慮して、潜在的な利益と潜在的な悪影響を文書化する。	
		システムの性質が原因で、ユーザーの属性に基づいてバイアスに関連する潜在的な被害が発生しやすいかどうかを評価する。	ユーザー属性には、人種、性別、年齢、障がい状態、その共通部分などが含まれる（これらに限定されない）。
	リスク軽減を文書化する	バイアスのリスクが存在する場合は、リスクを軽減するための取り組みを文書化する。	

設計

機能	カテゴリー	診断ステートメント	実装に関するコメント
プロジェクトの概念			
影響評価 *（続き）*	リスク軽減を文書化する	特定されたリスクと各リスクの潜在的な危険性をどのように測定するか、および軽減戦略の効果をどのように評価するかを文書化する。	
		バイアスのリスクが存在する場合は、リスクを軽減するための取り組みを文書化する。	
		リスクが軽減されていない場合は、リスクが許容可能と判断された理由を文書化する。	
リスク軽減のベスト・プラクティス	独立性と多様性	さまざまな利害関係者からフィードバックを得て、影響評価を行う。	この初期段階で特定されたリスクは、開発および影響評価プロセスの後半の側面を示すので、さまざまな経験、文化的背景、および専門知識を持つ人々から多様な視点を求めることで生じる可能性のある潜在的な被害について、総合的に理解することが決定的に重要である。社内の担当者が主題や文化的多様性を欠いている場合、第三者の専門家に相談するか、システムにより悪影響を及ぼす可能性のあるコミュニティのメンバーからフィードバックを求める必要がある。
	透明性の文書化	AI開発過程の後の段階で作業している担当者と影響評価の文書化を共有し、開発プロセス全体でリスクと潜在的な予想外の影響を監視できるようにする。	
	説明責任とガバナンス	潜在的なハイリスクAIシステムについて、上級管理職が十分に説明を受けていることを確認する。	「高リスク」と見なされるシステムの影響評価の文書化は、上級管理職と共有して、「実行 / 実行しない」意思決定を促進する必要がある。
データ取得			
影響評価	データの来歴の記録を維持する	AIモデルの訓練に使用されるデータの「再作成」を可能にするための十分な記録を維持し、その結果が再現可能であることを確認する。また、データソースに対する材料の更新を監視する。	記録には次のものを含める必要がある。 • データのソース • データの由来（作成者は誰か。時期はいつか。目的は何か。どのように作成されたか。） • データおよびデータガバナンス規則の使用目的および/または制限データを所有する法人はどこか。どのくらいの期間保管するか（または破棄する必要があるか）。使用に制限はあるか。 • 判明しているデータの制限（エレメントの欠落など） • データがサンプリングされた場合、どのようなサンプリング方針で行われたか。 • データは更新されるか。その場合、バージョンは記録されるか。

設計

機能	カテゴリー	診断ステートメント	実装に関するコメント
データ取得			
影響評価 *(続き)*	データに潜在的なバイアスがないか確認する	データを精査して過去のバイアスを確認する。	データのソースを確認し、過去のバイアスが含まれる可能性について評価する。
		データの「典型性」を評価する。	• 訓練データにおける人口分布を、システムを導入する集団のそれと比較する。 • システムを利用する可能性のある下位集団について説明するデータが十分にあるかを評価する。
		データのラベリング方法を精査する。	• データのラベル付けを行う担当者とプロセスを文書化する。 • 第三者データについては、潜在的なバイアス原因についてのラベルリング（および関連の方法）を精査する。
	リスク軽減を文書化する	バイアスを軽減するために、データが増強、操作、またはリバランスされたか、またその方法を文書化する。	
リスク軽減のベスト・プラクティス	独立性と多様性	データセットの細かな調査を容易にするため、データレビューチームにはそれぞれの分野における専門性や経験の観点において、多種多様な担当者を含める必要がある。	データにおいてバイアスの原因となりうる問題を効果的に特定するには、幅広い専門性や経験が必要になる（データが引用される領域について熟知していること、歴史的な背景や、その原因となった制度についての深い知識を持っていることなど）。社内担当者が多様性に欠ける場合、第三者の専門家または影響を受ける可能性のある利害関係者グループとの協議が必要になる可能性もある。
	非典型データのリバランス	追加データによるリバランスを検討する。	状況によっては、追加データを収集して全体的な訓練データセットのバランスを改善することで、代表性を向上させることができる。
		合成データによるリバランスを検討する。	不均衡なデータセットは、説明が不十分なグループのデータを「オーバーサンプリング」することで、リバランスできる可能性がある。一般的なオーバーサンプリング手法は、合成マイノリティオーバーサンプリング法(Synthetic Minority Oversampling Technique/SMOTE)である。この手法では、説明が不十分なグループから新しい「合成」データが生成される。

設計

機能	カテゴリー	診断ステートメント	実装に関するコメント
データ取得			
リスク軽減のベスト・プラクティス *（続き）*	データのラベリング	客観的で数値評価が可能なラベリング方針を定める。	• ラベリングバイアスの可能性を軽減するため、データのラベリング担当者には、個々のラベリング判断のための客観的で反復可能なプロセスを定めた明確な方針を与える必要がある。 • バイアスのリスクが高い領域では、ラベリング担当者が十分な専門知識を持ち、潜在的な無意識のバイアスを認識するためのトレーニングを受ける必要がある。 • ハイリスクなシステムでは場合によって、ラベリングの精度を監視するための品質保証の仕組みを設ける必要がある。
	説明責任とガバナンス	データラベリングプロセスを包括的なデータ戦略に統合する。	組織のデータ戦略を確立することで、データ評価を一貫して行うことができる。また、データを精査するための企業の取り組みが将来の参照用に確実に文書化されることで、作業の重複を防ぐことができる。

設計：リスク軽減ツールと参照資料

プロジェクトの概念

- ***Aequitas Bias and Fairness Audit Toolkit（Aequitasバイアスおよび公平性監査ツールキット）***
  Pedro Saleiro、Abby Stevens、Ari Anisfeld、およびRayid Ghani、シカゴ大学データ科学・公共政策センター（2018年）、http://www.datasciencepublicpolicy.org/projects/aequitas/）
- ***Diverse Voices Project | A How-To Guide for Facilitating Inclusivenss in Tech Policy（多様な声プロジェクト| 包括性を促進する技術ポリシーのハウツーガイド）***
  Lassana Magassa、Meg YoungおよびBatya Friedman、ワシントン大学技術政策研究所、https://techpolicylab.uw.edu/project/diverse-voices/

データ編集

- ***Datasheets for Datasets（データセットのためのデータシート）***
  Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé III、Kate Crawford、arXiv：1803.09010v7、（2020年3月19日）、https://arxiv.org/abs/1803.09010
- ***AI FactSheets 360（AIファクトシート360）***
  IBM Research、https://aif360.mybluemix.net/

開発

機能	カテゴリー	診断ステートメント	実装に関するコメント
データ準備とモデルの定義			
影響評価	特徴選択およびエンジニアリングプロセスを文書化する	特徴選択およびエンジニアリングプロセス中に行われた選択の根拠を文書化し、モデルの性能への影響を評価する。	特徴選択またはエンジニアリング選択が、暗黙のバイアスがある仮定に基づくかどうかを確認する。
		選択した特徴とセンシティブな人口属性との潜在的な相関関係を文書化する。	センシティブなクラスに密接に相関する特徴については、目標変数との関連性とそれをモデルに含める理由を文書化する。
	モデル選択プロセスを文書化する	選択したモデルアプローチの根拠を文書化する。	
		選択したアプローチの仮定とその結果生じる可能性のある制限を特定し、文書化し、および正当化する。	
リスク軽減のベスト・プラクティス	特徴選択	バイアスがあるプロキシ特徴を確認する。	• センシティブな属性をシステムへの入力として使用することを避けるだけでは(「無自覚の公平性」として知られるアプローチ)、バイアスのリスクを軽減する効果的なアプローチとはいえない。センシティブな特性がモデルから明らかに除外されている場合でも、他の変数がこれらの特性のプロキシとして機能でき、システムにバイアスをもたらす。プロキシバイアスのリスクを回避するために、AI開発者はモデルの特徴と保護された特性の潜在的な関係を確認し、これらのプロキシ変数がモデルの出力で果たす役割を確認する必要がある。 • AI開発者がセンシティブな属性データにアクセスできない場合および/またはそのようなデータを推測することが禁止されている場合は、特徴とセンシティブ属性との間の統計的な関係を確認する能力が制限される場合がある。[22]このような状況では、ドメイン専門家によるより包括的な分析が必要になる場合がある。

開発

機能	カテゴリー	診断ステートメント	実装に関するコメント
データ準備とモデルの定義			
リスク軽減のベスト・プラクティス *（続き）*	特徴選択	センシティブ属性に相関する特徴を詳細に調べる。	• センシティブ属性に相関することが知られている特徴は、システムの目標変数と強い論理的な関係がある場合にのみ使用すべきである。 • たとえば、所得は性別と相関しているが、ローンの返済能力とも合理的に関連している。したがって、信用力の評価を目的としたAIシステムで所得を用いることは正当化される。一方、信用力を予測するモデルで、同じく性別と関連する「靴のサイズ」を使用することは、センシティブ特性に密接に相関する変数を不適切に使用することになる。
	独立性と多様性	ドメイン固有の専門知識を持つ多様な関係者からのフィードバックを求める。	特徴エンジニアリングプロセスは、システムのトレーニングに使用されるデータの歴史的、法的、社会的側面について、多様な経験と専門知識を持つ担当者によって情報を与えられる必要がある。
	モデル選択	バイアスのリスクと潜在的影響の両方が高い状況では、不明瞭なモデルを避ける。	より解析しやすいモデルを使用すると、問題の特定と削減が容易になり、意図しないバイアスのリスクが軽減される。
モデルの検証、テスト、および改良			
影響評価	検証プロセスを文書化する	システム（および個々の要素）がどのように検証され、設計目標および意図する導入シナリオと一致して実行されているかを評価する方法を文書化する。	
		再検証プロセスを文書化する。	• モデルの定期的な再検証の周期を確立する。 • 周期外の再検証を行う性能基準を確立する。
	テストプロセスを文書化する	モデルの性能を評価して文書化することで、バイアスがないかシステムをテストする。	テストでは、設計フェーズで特定された公平性の数値基準を取り入れ、モデルの正確性とエラー率を属性グループ全体で確認する必要がある。
		テストの実施方法、公平性の数値基準の評価方法、およびその評価基準が選択された理由を文書化する。	
		モデル介入を文書化する	テストによって許容できないレベルのバイアスが明らかになった場合は、モデルを改良するための取り組みを文書化する。

開発

機能	カテゴリー	診断ステートメント	実装に関するコメント
モデルの検証、テスト、および改良			
リスク軽減のベスト・プラクティス	モデル介入	テスト中に表面化したバイアスに対処するために、モデルの改良の可能性を評価する。	選択した公平性の数値基準に基づいて、システムが許容できないレベルのバイアスを示していることがテストによって明らかになった場合は、モデルを改良する必要がある。次のようなモデルの改良が可能である。 • **処理前介入：**このような改良には、設計および開発ライフサイクルの初期段階を見直すことが含まれる場合がある（たとえば、追加の訓練データを探すこと）。 • **処理中介入：**バイアスは、モデルに追加の公平性の制約を直接与えることで軽減することもできる。従来の機械学習モデルは、予測精度を最大化するように設計されている。開発者は、新しい手法を使用してモデルに制約を構築し、グループ間のバイアスの可能性を低減することができる。公平性の制約を追加すれば、実際に、正確性と特定の公平性の数値基準の両方を最適化するようにモデルに指示できる。 • **処理後介入：**場合によっては、目的の分布に従わせるため、モデルの出力予測を操作する処理後アルゴリズムを使用してバイアスに対処できる。
	独立性と多様性	検証およびテストの文書化は、システムの開発に関与しなかった担当者が見直す必要がある。	独立したチームが、検証とテストの結果を、設計および開発プロセスの初期段階で作成したシステム仕様と比較する必要がある。

開発：リスク軽減ツールと参照資料

- ***Model Cards for Model Reporting（モデルレポートのためのモデルカード）***
  Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji、およびTimnit Gebru、2019年公平性・説明責任・透明性に関する会議議事録、（2019年1月）：220–229 、https://arxiv.org/abs/1810.03993
- ***AI Factsheets 360（AIファクトシート360）***
  Aleksandra Mojsilovic、IBM Research（2018年8月22日）、https://www.ibm.com/blogs/research/2018/08/factsheets-ai/
- ***AI Explainability 360（AI説明可能性360）***
  IBM Research、https://aix360.mybluemix.net/
- ***AI Fairness 360（AI公平性360）***
  IBM Research、https://aif360.mybluemix.net/
- ***Responsible Machine Learning with Error Analysis（エラー分析による責任ある機械学習）***
  Besmira Nushi、Microsoft Research（2021年2月18日）、https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774
- ***Aequitas Open Source Bias Audit Toolkit（Aequitasオープンソースバイアス監査ツールキット）***
  Pedro Saleiro、Abby Stevens、Ari Anisfeld、およびRayid Ghani、シカゴ大学データ科学・公共政策センター、（http://www.datasciencepublicpolicy.org/projects/aequitas/ ）
- ***FairTest: Discovering Unwarranted Associations in Data-Driven Applications（公平性テスト：データ駆動アプリケーションの不当保証の関連性発見）***
  FlorianTramer、Vaggelis Atlidakis、Roxana Geambasu、Daniel Hsu、Jean-Pierre Hubaux、Mathias Humbert、Ari JuelsおよびHuang Lin、arXiv、（2015年）、https://github.com/columbia/fairtest。
- ***Bayesian Improved Surname Geocoding（ベイズ推定による改良名字ジオコーディング）***
  消費者金融保護局（2014年）、https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

導入と使用

機能	カテゴリー	診断ステートメント	実装に関するコメント
導入と使用の準備			
影響評価	責任体系を文書化する	システムの出力とその結果を担当する担当者を定義し文書化する。これには必要に応じてシステムの決定を見直す方法の詳細も含まれる。	
		インシデントの可能性やシステムエラーの報告に対応するための管理計画を確立する。	• システム障害が発生するとどうなり、障害によって誰が被害を受ける可能性があるか。 • 障害はどのように検出されるか。 • 障害が検出された場合、誰が障害に対応しますか。 • システムを安全に無効にできるか。 • 重要な機能を継続するための適切な計画はあるか。
	データ監視のプロセスを文書化する	生産データ（システムの導入時に検出される入力データ）が訓練データと大きく異なるかどうかを評価するために用いるプロセスと数値基準を文書化する。	
	モデル性能を監視するプロセスを文書化する	静的モデルの場合は、性能のレベルとエラーの種類を時間の経過とともに監視する方法と、見直しを行う基準を文書化する。	
		時間の経過とともに変化するモデルの場合は、変更の一覧化方法、バージョンのキャプチャと管理方法、性能レベルの監視方法（スケジュール化された見直しの周期、周期外見直しを行う性能指標など）を文書化する。	
	監査プロセスと生産終了プロセスを文書化する	影響評価の監査を受ける周期を文書化し、リスク軽減制御が目的に適合しているかどうかを評価する。	
		システムサポートが提供される予定のスケジュールと、妥当な性能しきい値を下回った場合にシステムを廃止するためのプロセスを文書化する。	
リスク軽減のベスト・プラクティス	ドリフトとモデル劣化の監視	導入中に検出された入力データをシステムの訓練データの統計的表現と比較して評価し、データドリフトの可能性を評価することができる（訓練データと導入データの間の重大な相違により、モデルの性能が低下する可能性がある）。	

導入と使用

機能	カテゴリー	診断ステートメント	実装に関するコメント
導入と使用の準備			
リスク軽減のベスト・プラクティス *(続き)*	製品の特徴とユーザーインターフェース	製品とユーザーインタフェース特徴を統合して、予測可能な意図しない使用のリスクを軽減する。たとえば、人間参加型の要件を強化するインタフェース、システムが誤用された場合に通知する警告などである。	
	システムの文書化	AI開発者は、システムの能力、仕様、制限事項、意図する用途に関する十分な文書を提供し、AI導入者が導入リスクに関する独立した影響評価を実施できるようにする必要がある。	AI開発者は、必要に応じて、AI導入者に独立した影響評価を実施することができるように技術環境を提供することもできます。
		エンドユーザー使用許諾契約書に、予見可能な誤使用を防止するための制限事項を定めた条件を取り入れることを検討する(エンドユーザーが利用規定に準拠することを保証する契約上の義務など)。	
		販売およびマーケティング資料は、システムの実際の能力と合っていることを保証するため、詳細に検討する必要がある。	
	AIユーザー研修	AIの導入者は、システムの能力と制限、および出力を評価してワークフローに統合する方法について、AIユーザーに研修を提供する必要があります。	AIシステムの人間参加型監視を効果的なリスク軽減手段にするには、AIユーザーは適切な情報と研修の提供を受けて、システムの動作状況を理解し、モデルの出力を理解する必要がある。
	インシデントへの対応およびフィードバック機構	AIの導入者は、フィードバック機構を維持して、AIユーザーと影響を受ける個人(システムとやり取りする可能性のある一般のメンバー)がシステムの運用に関する懸念を報告できるようにする必要がある。	結果の決定においては、影響を受ける個人に異議申し立ての手段を提供する必要がある。

導入と使用:リスク軽減ツールと参照資料

- ***AI Incident Response Checklist (AIインシデント対応チェックリスト)***
  BNH.AI、https://www.bnh.ai/public-resources
- ***Watson OpenScale***
  IBM、https://www.ibm.com/cloud/watson-openscale
- ***Detect Data Drift on Datasets (データセットにおけるデータドリフトの検出)***
  Microsoft Azure Machine Learning(2020年6月25日)、https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python#create-dataset-monitors

基本資料

A Framework for Understanding Unintended Consequences of Machine Learning（機械学習の意図しない結果を理解するためのフレームワーク）
Harini SureshおよびJohn V. Guttag、arXiv（2020年2月）、https://arxiv.org/abs/1901.10002

AI Fairness（AI公平性）
Trisha Mahoney、Kush R.Varshney、Michael Hind、O'Reilly（2020 年4月）、https://www.oreilly.com/library/view/ai-fairness/9781492077664/

Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models（説明可能性を超えて：機械学習モデルにおけるリスク管理の実践ガイド）
Andrew Burt、Brenda Leong、Stuart Shirrell、およびXiangnong（George）Wang、プライバシーの未来フォーラム（2018年6月）、https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI（AIの公平性に関する組織における課題と機会を理解するための参加型デザインチェックリスト）
Michael A. Madaio、Luke Stark、Jennifer Wortman Vaughan、Hanna Wallach、CHI2020：2020年CHI人と情報システムの相互作用に関する国際会議議事録（2020年4月）、1–14、https://doi.org/10.1145/ 3313831.3376445

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing（AI説明責任のギャップを埋める：内部アルゴリズム監査のためのエンドツーエンド・フレームワークの定義）
Raji I.D.、Smart A.、White R. N.、Mitchell M.、Gebru T.、Hutchinson B.、Smith-Loud J.、Theron D.、およびBarnes P.、FAT* 2020：2020年公平性・説明責任・透明性に関する会議議事録、（2020年1月）：33–44、https://doi.org/10.1145/3351095.3372873

Supervisory Guidance on Model Risk Management（モデルリスク管理に関する監督ガイダンス）
米国連邦準備制度理事会（2011年4月）、https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector（人工知能の倫理と安全性の理解：民間セクターにおけるAIシステムの責任ある設計および実施ガイド）
David Leslie、アラン・チューリング研究所（2019 年）、https://doi.org/10.5281/zenodo.3240529

注釈

[1] Gina Kolata、"Alzheimer's Prediction May Be Found in Writing Tests," (「アルツハイマーの予測は筆記試験で見出される」)　*ニューヨークタイムズ*(2021年2月1日)、https://www.nytimes.com/2021/02/01/health/alzheimers-prediction-speech.html

[2] Dina Temple-Raston、*Elephants under Attack Have an Unlikely Ally: Artificial Intelligence (攻撃を受けた象には意外な味方がある：人工知能)*、NPR(2019年10月25日)、https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence

[3] *Seeing AI: An App for Visually Impaired People That Narrates the World Around You (Seeing AI：まわりの世界を語る視覚障がい者のためのアプリ)*、Microsoft、https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/

[4] たとえば、Jennifer Sukis、*The Origins of Bias and How AI May Be the Answer to Ending Its Reign (バイアスの起源およびその支配を終わらせる方法の答えがAIであることについて)*、Medium(2019年1月13日) https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354を参照のこと。

[5] たとえば、Nicol Turner Lee、Paul Resnick、およびGenie Barton、*Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms (アルゴリズムのバイアスの検出と緩和：消費者被害を低減するためのベスト・プラクティスと方針)*、ブルッキングス(2019年5月22日) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/を参照のこと。

[6] Harini Suresh およびJohn V. Guttag、*Understanding Unintended Consequences of Machine Learning (機械学習の意図しない結果を理解するためのフレームワーク)*(2020年2月17日)、https://arxiv.org/pdf/1901.10002.pdf

[7] Xiaolin WuおよびXi Zhang、*Automated Inference on Criminality Using Face Images (顔画像を用いた犯罪者の自動推論)*、上海交通大学(2016年11月13日)、https://arxiv.org/pdf/1611.04135v1.pdfを参照のこと。

[8] Blaise Aguera y Arcas、Margaret Mitchell、Alexander Todorov、*Physiognomy's New Clothes (観相学の新しい様相)*、Medium(2017年5月6日)、https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[9] *Ziad Obermeyer、Brian Powers、Christine Vogeli、Sendhil Mullainathan、"Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," (「公衆衛生を管理するために使用されるアルゴリズムにおける人種バイアスの分析」)、サイエンス (2019年10月25日)*、https://science.sciencemag.org/content/366/6464/447

[10] Solon BarocasおよびAndrew D. Selbst、"Big Data's Disparate Impact," (「ビッグデータの異なる効果」) 、*カリフォルニア大学ロー・レビュー*104、no.3(2016年9月30日)：671 、http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

[11] Joy BuolamwiniおよびTimnit Gebru、"Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," (「ジェンダー・シェード：商業上のジェンダー分類における交差的正確性格差」)*Machine Learning Research 議事録*81(2018年)：77-91、http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[12] Kate Crawford、*The Hidden Biases in Big Data (ビッグデータの隠れたバイアス)*、ハーバード・ビジネス・レビュー(2013年4月1日)、https://hbr.org/2013/04/the-hidden-biases-in-big-data

[13] Kate CrawfordおよびTrevor Paglen、*Excavating AI: The Politics of Images in Machine Learning Training Sets (AIの発掘：機械学習トレーニングセットの画像の政治学)*2019年9月19日)、https://excavating.ai/

[14] Cade Metz、「"'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?" (「『オタク』、『禁煙者』、『非行者』：A.Iはあなたにどんなレッテルを貼るか？」)*ニューヨークタイムズ*(2019年9月20日)、https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html

[15] Jessica Zosa Forde、A. Feder Cooper、Kweku Kwegyir-Aggrey、Chris De Sa、Michael Littman、*Model Selection's Disparate Impact in Real-World Deep Learning Applications (現実世界におけるディープラーニング・アプリケーションに対するモデル選択の異なる効果)*、arXiv：2104.00606(2021年4月1日)、https://arxiv.org/abs/2104.00606

[16] Aaron Klein、*Credit Denial in the Age of AI (AI時代の融資拒絶)*、ブルッキングス研究所(2019年4月11日) https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/

[17] J. Vaughn、A. Baral、M. Vadari "Analyzing the Dangers of Dataset Bias in Diagnostic AI systems: Setting Guidelines for Dataset Collection and Usage," (「 診断AIシステムにおけるデータセットバイアスの危険性の分析：データセットの収集と使用のガイドラインの設定」)、ACM健康・推論と学習に関する会議、2020年ワークショップ、http://juliev42.github.io/files/CHIL_paper_bias.pdf

[18] Arvind Narayanan、*21 Fairness Definitions and Their Politics (2021年公平性の定義とその政治学)、ACM公平性・説明責任・透明性に関する会議*(2018年3月1日)、https://www.youtube.com/watch?v=jIXIuYdnyyk

[19] Reuben BinnsおよびValeria Gallo、*AI Blog: Trade-Offs (AIブログ：トレードオフ)*、英国個人情報保護監督機関(2019年7月25日)、https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/

[20] Inioluwa Deborah Rajiら、*Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing (AI説明責任のギャップを埋める：内部アルゴリズム監査のためのエンドツーエンド・フレームワークの定義)、FAT* 2020：2020年公平性・説明責任・透明性に関する会議議事録*(2020年1月)：33–44、https://doi.org/10.1145/3351095.3372873

[21] Sara Hooker、"Algorithmic Bias Is a Data Problem," (「アルゴリズムバイアスはデータ問題である」を乗り越えて)、*Patterns*、(2021年4月9日) 、https://www.sciencedirect.com/science/article/pii/S2666389921000611

[22] McKane Andrus、Elena Spitzer、Jeffrey Brown、Alice Xiang、*"What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness (「測定できないことは、理解できないこと」：公平性の追求におけるデモグラフィックデータ調達の課題)*、arXiv：2011.02282(2021年1月23 日)、https://arxiv.org/abs/2011.02282